Gac Sanit. 2014;28(5):408–410

View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Nota metodológica

Uso equitativo de tests en ciencias de la salud

Albert Espelt[a,b,c,*], Carme Viladrich[b], Eduardo Doval[b,d], Joan Aliaga[b], Rebeca García-Rueda[b] y Salomé Tárrega[b]

[a] Agència de Salut Pública de Barcelona, Barcelona, España
[b] Departament de Psicobiologia i de Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), España
[c] CIBER de Epidemiologia y Salud Pública (CIBERESP), España
[d] Grupo de Investigación en Estrés y Salud (GIES), Universitat Autònoma de Barcelona, Bellaterra (Barcelona), España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 14 de marzo de 2014
Aceptado el 3 de mayo de 2014
On-line el 10 de junio de 2014

Palabras clave:
Sesgo
Psicometría
Reproducibilidad de resultados
Equidad

RESUMEN

Los instrumentos de medida estandarizados (tests) se han convertido en una herramienta imprescindible en las ciencias de la salud. En los estándares para pruebas educativas y psicológicas publicados en 1999 por la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education se introduce el concepto de equidad en el desarrollo, la adaptación y la administración de los tests psicométricos. A pesar de su innegable relevancia, este concepto ha sido poco utilizado en el mundo de la salud pública y la epidemiología. Por ello, la presente nota metodológica tiene como objetivo explicar el concepto de equidad en los tests y dar herramientas e indicaciones para detectar y solventar su uso no equitativo.

© 2014 SESPAS. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

Fair use of tests in health sciences

Keywords:
Bias
Psychometrics
Reproducibility of results
Fairness

ABSTRACT

Standardized measurement instruments (tests) have become an essential tool in health sciences. The concept of equity in the development, adaptation and administration of psychometric tests was first introduced in «Standards for Educational and Psychological Testing» published in 1999 by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Despite its importance, this concept has been scarcely used in epidemiology and public health. Consequently, this methodological note aims to explain the concept of equity in testing and to provide tools and indications to detect and solve their inequitable use.

© 2014 SESPAS. Published by Elsevier España, S.L.U. All rights reserved.

Introducción al concepto de test equitativo

Los investigadores y otros profesionales que utilicen tests deben garantizar que las interpretaciones de las puntuaciones obtenidas sean válidas para la finalidad planteada, que sean también fiables y que, además, tanto su proceso de administración como los resultados obtenidos y sus consecuencias sean equitativos[1,2]. Un test es equitativo[1,2] si permite evaluar a las personas de manera imparcial, es decir, sin que aspectos no relevantes para lo que se pretende medir con el test, como por ejemplo el sexo o la etnia, tengan una influencia destacada sobre los resultados de la evaluación. Podría parecer que la equidad se garantiza asegurando que todas las personas reciban el mismo trato durante el procedimiento de evaluación. De hecho, la mayoría de los tests se aplican e interpretan siguiendo un protocolo estandarizado[1]. Sin embargo, paradójicamente, en algunas ocasiones estas condiciones igualitarias pueden generar desigualdad y producir sesgos contra determinados grupos de personas. Por ejemplo, a principios del siglo xx, se observaron diferencias en los tests de inteligencia aplicados a la población norteamericana y a los inmigrantes que llegaban al país. Aunque los tests aplicados y sus condiciones de administración eran iguales para ambos grupos, las diferencias culturales y el poco dominio del idioma perjudicaban a muchos inmigrantes cuyas puntuaciones resultaban claramente sesgadas, y por tanto también las conclusiones que se derivaban de ellas[2].

Consecuencias de la inequidad

De un test no equitativo pueden derivarse graves consecuencias tanto individuales (p.ej., dos personas con el mismo problema de salud podrían ser diagnosticadas de manera diferente según su sexo) como grupales (p.ej., una misma intervención podría dar resultados distintos en diferentes grupos sociales), de manera que es importante conocer las causas frecuentes de inequidad, las formas de detectarla y las acciones a tomar para convertir un test en equitativo.

La mayoría de las barreras para la equidad están relacionadas con el formato del test y con aspectos de su administración. Por

* Autor para correspondencia.
Correo electrónico: aespelt@aspb.cat (A. Espelt).

http://dx.doi.org/10.1016/j.gaceta.2014.05.001
0213-9111/© 2014 SESPAS. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

ejemplo, el redactado de algunas preguntas o el tamaño de la letra pueden suponer un problema de comprensión para personas sin un buen dominio del idioma o con dificultades visuales[3]. En otras ocasiones, las barreras tienen su origen en la estructura y el diseño del test. Esto ocurre cuando los aspectos evaluados no son compartidos por toda la población a la cual va dirigido el test[4] o cuando no hay invariancia métrica en los resultados de la prueba[5].

Identificación de la inequidad

Aunque la evaluación y la corrección de cada tipo de barrera afectan a el proceso de desarrollo, adaptación y administración de un test, como mínimo debería comprobarse que en los resultados finales no se observan ítems sesgados. Un ítem está sesgado cuando presenta un funcionamiento diferencial (DIF, *differential item functioning*)[6] que no se explica por un posible impacto. Un ítem presenta DIF cuando la respuesta de una persona no sólo depende de su nivel en la característica evaluada, sino también de su grupo de pertenencia (p.ej., sexo o clase social). Existen varias técnicas para identificar DIF y su elección suele fundamentarse en criterios tan variados como la teoría psicométrica de base, el tipo de ítems de la prueba, el criterio de referencia, el tamaño muestral de los grupos o la simplicidad computacional[2,7,8]. En cualquier caso se modelizan las respuestas al ítem como consecuencia de dos predictores: el primero, un criterio interno o externo de referencia para lo que se pretende medir[7], y el segundo, el grupo de pertenencia que se sospecha que puede generar diferencias. En el ámbito de la salud pública, una solución sencilla sería utilizar las puntuaciones totales del test como criterio, y la regresión logística como técnica de análisis. En este caso, la probabilidad de dar una respuesta al ítem *(P(u = 1))* en un conjunto de *n* individuos *(i)* vendría expresada por $P(u_i = 1) = \frac{e^{z_i}}{(1+e^{z_i})}$, siendo $z_i = \beta_0 + \beta_1\theta_i + \beta_2 g_i + \beta_3(\theta_i g_i)$, θ_i la puntuación total al test y g_i la pertenencia a uno de los grupos. Podemos evaluar si, con independencia de la puntuación del test:

1) No existe DIF $[z_i = \beta_0 + \beta_1\theta_i]$, es decir, si sólo el coeficiente asociado a la variable puntuación total del test es estadísticamente distinto de cero.
2) Existe DIF uniforme, es decir, siempre a favor del mismo grupo $[z_i = \beta_0 + \beta_1\theta_i + \beta_2 g_i]$, si además también es estadísticamente distinto de cero el coeficiente asociado al grupo de pertenencia.
3) Existe DIF no uniforme, es decir, hay interacción apreciable entre el criterio y la pertenencia al grupo $[z_i = \beta_0 + \beta_1\theta_i + \beta_2 g_i + \beta_3(\theta_i g_i)]$, si además el coeficiente asociado a la interacción de la puntuación total y el grupo de pertenencia es estadísticamente distinto de cero.

Las puntuaciones simuladas del ítem «¿te has sentido preocupado/a por tu consumo de alcohol?» de la *Severity of Dependence Scale* (SDS)[9] (fig. 1) muestran un DIF uniforme, ya que para cada nivel de dependencia al alcohol (puntuación total en la escala SDS) las mujeres puntúan más alto que los hombres en dicho ítem. Este resultado simulado podría ser un indicador de sesgo si las mujeres se sintiesen preocupadas no sólo por su dependencia al alcohol sino por la presión social añadida a la que pueden estar sometidas las mujeres con respecto al consumo del alcohol, un aspecto que puede desligarse perfectamente de la dependencia al consumo. En la medida en que se deriven consecuencias del resultado de este test, el ítem trataría de forma inequitativa a hombres y mujeres.

Ésta es precisamente la segunda condición necesaria para concluir que el ítem está sesgado, y consiste en descartar diferencias reales en el criterio entre los grupos evaluados. Cuando las diferencias son reales decimos que se trata de impacto. Por ejemplo, al comparar mayores y jóvenes en la respuesta al ítem de función física del SF36 «Su salud actual, ¿le limita para hacer esfuerzos intensos,

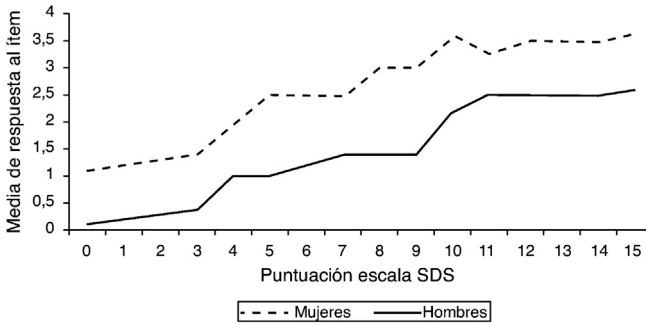

Figura 1. Simulación de la relación entre la puntuación media de respuesta al ítem «¿te has sentido preocupado por tu consumo de alcohol?» y la puntuación total de la *Severity of Dependence Scale* (SDS) en hombres y mujeres.

tales como correr, levantar objetos pesados o participar en deportes agotadores?» se obtiene un DIF uniforme a favor de las personas jóvenes, pero es explicable por el impacto real de la función física de ambos grupos de edad sobre dicho ítem. La presencia de impacto no tiene connotaciones negativas para el test.

Para determinar si el DIF es atribuible a sesgo o a impacto, decidir qué hacer con los ítems sesgados y seleccionar medidas correctoras[10], deberían utilizarse técnicas de investigación cualitativas[1] (p.ej., panel de expertos, personas afectadas).

Prevención de la inequidad

Una de las medidas correctoras más utilizadas para prevenir la inequidad es el uso de acomodaciones que pueden diseñarse o aplicarse en las fases de desarrollo, adaptación o administración de un test a poblaciones específicas. La acomodación hace referencia a cualquier acción que modifique el protocolo establecido en un test, ya sea en los contenidos o en su administración, para aplicarlo a una persona o grupos de personas, sin afectar por ello al constructo que se pretende medir[1,7]. Por ejemplo, podríamos salvar el problema de inequidad en la SDS acomodando la redacción de las instrucciones con unas específicas dirigidas a las mujeres para intentar que contesten al ítem a pesar de los condicionantes sociales. Las modificaciones del test pueden hacerse cambiando la presentación, el formato de respuesta, el lugar de aplicación o el tiempo de respuesta, o bien administrando sólo una parte del test o realizando evaluaciones alternativas. Sin embargo, hay que tener precaución al elegir el tipo de acomodación porque en ocasiones puede ser inadecuada[7]. También hay que tener cautela al interpretar las puntuaciones obtenidas después de acomodar contenidos o procedimientos de administración de un test.

Por último, queremos señalar que la documentación de un test debería incluir información sobre la validez y la fiabilidad de las puntuaciones, así como criterios interpretativos para cada uno de los subgrupos que tenga sentido comparar y también las comprobaciones relativas a la influencia de la acomodación sobre los resultados psicométricos. Sólo de esta manera los profesionales podrán tomar decisiones responsables sobre los tests que utilizarán.

Editora responsable del artículo

Mª Felicitas Domínguez-Berjón.

Contribuciones de autoría

A. Espelt y C. Viladrich diseñaron el trabajo. A. Espelt, C. Viladrich y E. Doval escribieron y discutieron la primera versión del manuscrito. A. Espelt, C. Viladrich, E. Doval, J. Aliaga, R. García-

Rueda y S. Tárrega, revisaron esta primera versión y contribuyeron a las sucesivas versiones. Todas las personas autoras del manuscrito revisaron y aprobaron su versión final.

Financiación

Ninguna.

Conflicto de intereses

Ninguno.

Bibliografía

1. American Educational Research Association. American Psychological Association, National Council on Measurement in Education. En: Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
2. Gómez-Benito J, Hidalgo MD, Guilera G. El sesgo de los instrumentos de medición. Tests justos. Papeles del Psicólogo. 2010;31:75–84.
3. Choi BCK, Pak AWP. A catalog of biases in questionnaires. Prev Chronic Dis. 2005;2:A13.
4. Rasmussen A, Katoni B, Keller AS, et al. Posttraumatic idioms of distress among Darfur refugees: Hozun and Majnun. Transcult Psychiatry. 2011;48: 392–415.
5. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ Res Methods. 2000;3:4–70.
6. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Quality of Life Research. 2007;16:33–42.
7. Abad F, Olea J, Ponsoda V, et al. Medición en ciencias sociales y de la salud. Madrid: Editorial Síntesis; 2011. p. 566.
8. Choi SW, Gibbons LE. Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw. 2011;39:1–30.
9. Cuenca-Royo AM, Sánchez-Niubó A, Forero CG, et al. Psychometric properties of the CAST and SDS scales in young adult cannabis users. Addict Behav. 2012;37:709–15.
10. Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes. 2010;8:1–9.